Using league table rankings in public policy formation: statistical issues.

Harvey Goldstein

University of Bristol

Abstract

This chapter reviews the statistical models that underpin institutional comparisons based upon outcome measures for their students. The strengths and limitations of inferences from these models are explored, with examples taken from education.

Keywords

League tables, institutional performance, multilevel models

Correspondence

Harvey Goldstein

Professor of Social Statistics

University of Bristol

Bristol, BS8 1JA

UK

h.goldstein@bristol.ac.uk

Introduction

One of the striking features of public policy over the last three decades has been the growing utilisation of quantitative measures of institutional performance in many countries to make judgements and allocate resources. The key feature of these systems is that the resulting rankings of institutions such as schools, hospitals and police forces are published by government bodies and publicised by the media. They are to be distinguished from 'intelligence systems' [1] that also produce rankings but rather than publishing these widely, use them to inform the institutions themselves and those responsible for monitoring them. Such systems operate, for example, in the area of public transport and have also been described in an educational context [2]. It has been argued that these latter systems have advantages over published systems [3] in that they minimise unwanted or 'perverse' side effects such as 'gaming' to improve ranking position. They also address directly the underlying issues of how institutional performance can be improved, rather than indirectly attempting to achieve this by exposing current performance for public scrutiny. While in many respects the statistical issues associated with design and analysis are the same, in this review I shall not consider intelligence systems in any detail, but concentrate on a discussion of public rankings. I shall deal largely with rankings in the area of education, especially school education. Health is dealt with separately in this volume [4] and the statistical issues in other areas are similar.

This review will address itself to the issues associated with the design of ranking systems, and the modelling and interpretation the results of published rankings. Ideally there would be a discussion of evaluations of the effects of such systems, but such evaluations, however desirable, are rare and typically not envisaged when systems are designed. Where attempted, however, these will be noted. The next section will set out some basic concepts, and this is followed by sections that address different areas of application and technicalities. This is followed by some examples and I will conclude with recommendations and areas for further work.

Constructing rankings

To illustrate the process of ranking construction consider the case of school education where data are available from individual students attending each school. These may be in the form of, for example, responses to a questionnaire seeking views about satisfaction with teaching, or the results of test or examination scores. Typically, data are chosen to represent a particular time point or period, such as the age at which external tests for admission to higher education are taken. At its simplest a ranking will be formed by calculating the mean value, over students, and

then producing a ranked list of these means. There may be several of these, for example, for different curriculum subjects and for some purposes these may be averaged into a single index using weights, as is done for rankings of universities [5]. This is an example where rankings are based upon *aggregate* measures made at the institutional level. For universities these would include such things as reputation among peers and measures of total research output. I will discuss issues associated with these in a later section. First, I will deal with cases where linked information is available on individuals within institutions as is often the case with schools.

Aggregate rankings have been subject to criticism on two broad counts. The first is that they fail to 'contextualise' the results by taking into account factors over which institutions have little control but which nevertheless have a strong association with the results. A particularly important factor is a selective intake so that for example, hospital units may have different case mixes with some having higher risk patients than others, and schools that recruit 'high achieving' students will be expected to have higher test and exam scores than others, irrespective of the quality of the schooling received. If that is the case then the ostensible purpose of the ranking, namely to compare schooling quality, will be undermined. A number of authors have discussed this issue and shown how such contextual factors can be incorporated as covariates in a model based 'value added' approach [6-8]. The following model captures the essence of such approaches and will be elaborated as appropriate.

The basic data structure is that of a 2-level hierarchy with students nested within schools, or patients nested within hospitals etc. The standard approach to describing such data is via a multilevel or random effects model as follows [9]

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_j + e_{ij}$$

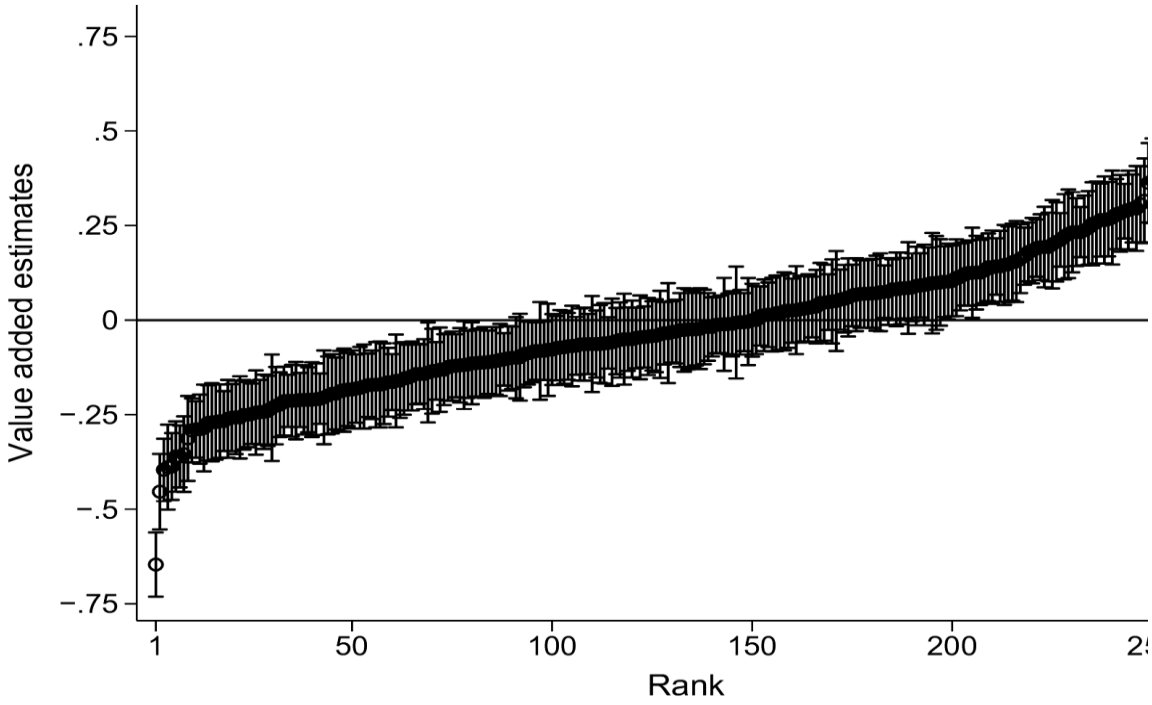
$$e_{ij} \sim N(0, \sigma_e^2), \quad u_j \sim N(0, \sigma_u^2) \qquad (1)$$

where for simplicity we assume mutually independent normal random effects. Our response, for example an examination score at the end of secondary (high) school, for student *i* in school *j* is y_{ij} and x_{ij} is, for example, a prior achievement test score designed to capture any selection by prior achievement. In the next section we will discuss this model further, but for now we shall describe how, with appropriate data, we can derive rankings.

Ignoring the possibility of any missing data model (1) is fitted to all those students in the sample, which may be the total number of pupils in each school year group or cohort. Under the normality assumption we can obtain parameter estimates and also posterior estimates for each of the random effects u_j. If maximum likelihood is used these effects are the usual shrunken residuals, and the equivalent posterior estimates can be obtained from a straightforward Bayesian analysis, for example using a Gibbs sampler with diffuse priors [9, Chapter 2]. The

following figure shows a ranking of a sample of 266 schools in England with a median cohort size of 190 [10].

Figure 1.



The response is a normalised examination score taken at the end of compulsory secondary schooling in grade 11 and this is adjusted using a test score taken prior to entry to secondary school, ethnic group and various measures of social and linguistic disadvantage. The vertical bars are conventional 95% normal intervals. It can be seen that about half of the schools have an interval that includes the population mean. If we wish to compare two chosen schools in terms of whether their intervals overlap, then an appropriate interval length for this can be computed and is approximately 0.7 times the intervals displayed above [11]. We also note that similar results can be displayed directly in terms of rankings, rather than residual estimates, with corresponding intervals [8].

Such displays can be used to provide basic information about school 'effects' that may be of use in terms of more detailed follow-up of individual institutions by, for example inspectors. They have also been advocated for use, for example, by students' parents as an aid in choosing schools, although their use as such is problematic since what is really required is a prediction of school effects several years ahead and this adds extra uncertainty that makes any statistical separation very difficult [10 and is further explored in a later example. Alternative formulations have been proposed based upon measuring the effects of individual teachers during a period of schooling and this will be dealt with below.

The use of such adjusted or 'value added' models is intuitively appealing as an improvement on the use of unadjusted means and has become reasonably well established in the area of schooling. In other areas, however, this approach is more problematic. In the case of the expanding area of university rankings (see for example [12]) it is typically difficult to find sensible adjustment variables, and measuring and linking

these together for individuals within universities would also be difficult. Furthermore, it is often not clear what the purpose of such rankings is. Thus, if the intention is to provide choice for undergraduate applicants, then a measure of degree outcome or satisfaction would seem appropriate, but such measures tend to be incomparable both across universities and across disciplines. If the intention is to provide overall measures of research performance there are further difficulties with defining this in terms of factors such as citation indices and there seems to be little consensus about this [13]. Another difficulty lies with the typical requirement to combine individual indicators into a single measure for presentational purposes and this will involve decisions about which weights to use in such a process. A more detailed discussion is given in [3] and I shall return to this issue later.

I now look at various extensions and practical issues and how these may be dealt with.

The implementation of adjusted ranking models

The first key area of concern is with the nature of the criterion being used to judge institutional performance. It is not only with university rankings that this issue is discussed. In the area of schooling, while the use of test scores is common, there are objections to this in terms of a resulting over-concentration by schools on improving test scores at the expense of broader educational measures or promoting some students at the expense of others in order to improve their league table position. Likewise in measuring aspects of policing the choice of outcome is debateable. While such debates are important they are additional to the technical concerns of this chapter and will not be pursued, but see [3].

When adjustments are incorporated into ranking models these typically are based upon measures taken at an earlier time or set of times. They are distinct from 'scaling' adjustments that might be used, for example, to measure university research output *per academic* where a measure of total output is scaled by an estimate of the number contributing to it. Such an estimate may not be straightforward to compute, but this is a measurement problem rather that a modelling one per se. On the other hand, a university drop-out rate might need to be adjusted for intake measures in order to avoid, for example, manipulation of results by exclusion of students from underprivileged backgrounds more likely to drop out for financial reasons. In the following sections I shall consider several relevant issues, including the adequacy of prior measures used, student mobility, differential school effects, missing data, endogeneity and other aspects of model misspecification.

Prior information

Most league table rankings, whether of schools or other institutions utilise a single measure of prior performance to adjust for selection factors, whether purposeful or haphazard. However,

there is evidence [14] that in the case of secondary schooling information about prior school attended and achievement during that period of school will generally change inferences, although other authors [15] suggest that in primary (elementary) schools rankings are relatively unaffected when a sequence of prior achievement measures is used as opposed to just one.

Moving across institutions

In practice many students will change school over the course of the period of schooling. Most rankings are published using the school membership at the time at which the outcome measure is taken. In most systems, however, there is movement during the course of a period of schooling so that the contributions of all schools attended should be taken into account. When this is done [15] the variation attributable to schools generally increases, but again seems to have little effect on the rankings.

To take account of such mobility we may use a multiple membership model. This involves extending (1) as follows. The following is a model for just two schools (1,2) for simplicity, where students can move between them.

$$y_{i\{j_1 j_2\}} = \beta_0 + \beta_1 x_{i\{j_1 j_2\}} + w_{1ij_1} u_{j_1} + w_{1ij_2} u_{j_2} + e_{i\{j_1 j_2\}}$$

$$w_{1ij_1} + w_{ij_2} = 1 \qquad (2)$$

$$e_{ij} \sim N(0, \sigma_e^2), \quad u_j \sim N(0, \sigma_u^2)$$

This states that the random effect contribution to the response is a weighted combination of the random effects associated with the schools attended. For several schools we will have contributions from one or more with associated weights. The weights have to be chosen, for example proportional to the time spent at each institution. We shall see later that in some cases these weights can be estimated. In practice we can carry out sensitivity analyses with different weighting functions, possibly choosing the one that produces the best 'fit', for example as judged by the DIC statistic in a Bayesian analysis [16]. One consequence of (2) is that the total level 2 variance has the form $\sigma_u^2 \sum_h w_{ih}^2 \leq \sigma_u^2$, so that ignoring mobility will lead to an underestimate of the level 2 variance. Further details on fitting such models are given in [9, Chapter 13].

Moving within institutions

In the case of schools, especially secondary or high schools, data may be available longitudinally for students at the end of each year of schooling, and thus attached to different teachers who will provide separate effects on outcome measures. Multiple membership models such as (2) can be adapted for this situation [6]. Such a model can be written as

$y_{tij} = (X\beta)_{tij} + u_{tj} + \sum_{t^*<t} \alpha_{tt^*} u_{t^*} + e_{tij}\,, \quad t = 1, \ldots p$ (3)

where at the end of year *t* we have a contribution from the current teacher (u_{tj}) and contributions from all the different teachers prior to year *t* (u_{t^*}). The covariates (*X*) can include prior attainment as well as, for example, socio-economic indicators and school level variables. This model is known as the general persistence model ($\alpha_{tt^*} < 1$) with a special case being the 'complete persistence' model where $\alpha_{tt^*} = 1$, indicating that each previous teacher has the same effect on a future outcome irrespective of how far ahead this may be. We also allow the level 1 (occasion) residuals for a student to be correlated across occasions. It is also assumed that there is enough movement among groups of students from year to year to enable identification of the model parameters.

To illustrate this model in a simple case with just 3 occasions, we have

$y_{1ij} = \beta_{10} + u_{1j} + e_{1ij}$

$y_{2ij} = \beta_{20} + u_{2j} + \alpha_{21} u_{1j} + e_{2ij}$

$y_{3ij} = \beta_{30} + u_{3j} + \alpha_{32} u_{2j} + \alpha_{31} u_{1j} + e_{2ij}$

This basic model can be extended in a number of ways.

- If we have several teachers for each student in any given year then we can introduce standard multiple membership weights for each student where these add to 1.0 as in (2).
- The pupil level residual covariance matrix can be structured as a function of time to reduce the number of parameters, for example $e_{tij} = e_{0i} + e_{1i} t + \delta_{tij}$.
- Extra levels, such as that of school, or cross classifications can be introduced.
- Generalised linear models can be used.
- We can accommodate the same teacher in more than one year by modifying the indicator matrix for the teacher random effects.
- A multivariate extension is possible whereby we can model outcomes in more than one curriculum subject.

It is possible that in any given dataset we may be missing the teacher identification for some students for some years. In this case one approach is to assume that all those students with a missing teacher identification in any given year belong to a new 'pseudo' teacher and sample accordingly. An alternative is to assume that the true teacher is one of those for whom data are available and use weights similar to multiple membership weights corresponding to the observed distribution of students among these teachers.

These models are used in many State education systems in the United States of America to evaluate teachers and an introduction to a series of papers discussing their strengths and weaknesses can be found in Beardsley et al. [20]. A key issue is that the confidence intervals

associated with any one teacher tend to be large and sensitive to the assumptions of the model [6].

Differential effectiveness

So far I have assumed a simple (random) effect for an institution. There is now a great deal of evidence, at least for schooling, that the institutional effect will also depend upon the characteristic of the individual, for example whether a girl or a boy or an initial low or high achiever [17]. This can be incorporated using a random coefficient model such as the following where we allow different random effects for boys and girls, parameterised in terms of an overall school effect and one for the boy-girl difference.

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + u_{0j} + u_{1j} x_{2ij} + e_{ij} \qquad (4)$$

or alternatively

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + u_{0j} + u_{2j} x_{2ij} + e_{ij}$$

$$\beta_{0j} = \beta_0 + u_{0j}, \quad \beta_{2j} = \beta_2 + u_{2j}$$

$$\begin{pmatrix} u_0 \\ u_2 \end{pmatrix} \sim N\left(\begin{matrix} 0 \\ 0' \end{matrix} \begin{matrix} \sigma_{u0}^2 \\ \sigma_{u02} \end{matrix} \quad \sigma_{u2}^2 \right), \qquad e_{ij} \sim N(0, \sigma_e^2)$$

When such a model is introduced comparisons may alter considerably. Thus Yang et al. [18] show that rankings of primary schools can be very different for initially high and low achievers. They also demonstrate that miss-specifying the model by ignoring a random coefficient for initial test score results in a spuriously high correlation between the adjusted and unadjusted school effects.

Endogeneity and model misspecification

Some forms of model misspecification can lead to endogeneity, whereby one or more covariates in the model is correlated with the random effects, in particular with the level 2 residuals. For example, in the model given by (4), if the term $u_{1j}x_{2ij}$ is omitted then this component of the level 2 variation will be absorbed into the u_{0j} with the result that these random effects will, in part, depend on the x_{2ij} so inducing a relationship with the covariate in the fixed effects part of the model. Such an omission will lead generally to misleading inferences for institutional rankings. In general it would seem to be preferable to report and base decisions on estimates from the full model, where these are available. It may be the case that, for some purposes, however, we would wish to estimate the parameters of the simpler model given by (1), effectively marginalising (4) over the random coefficient effects. By default, such marginalisation will typically be carried out with respect to the observed sample, assumed to be representative of a suitably defined population. In this case it is easy to show that a standard estimation, for

example using maximum likelihood, that assumes (1) is the correct model, will provide biased estimates for the required marginal model and this is an example of a failure to take account of endogeneity.

The issue, however, is not straightforward since we may choose to marginalise with respect to a different population structure. For example, it would seem reasonable in some circumstances to 'standardise' the within-school distribution of gender, in order to adjust for different proportions of male and female students which are considered irrelevant to the making of comparisons. In such a case we might choose to marginalise with respect to having equal numbers within each school. We can think of this as applying equal weights to the male and female effects for each school. Another example would be where the coefficient of prior achievement varied randomly across schools and marginal estimates might be desired for a 'standard' distribution of prior achievement so that schools could be compared directly having after prior achievement was adjusted for. To carry out such marginalisations, for example using bootstrap methods, we first need to fit the fully specified model. Thus, alternative methods such as GEE (see for example Hubbard et al. [19]) that fit marginal methods directly, effectively using the observed sample structure, are generally not appropriate.

The issue of endogeneity is really part of a general concern with model misspecification, rather than the narrower concern with biased estimates for a particular marginal model.

Measurement error

Measurement or category misclassification errors will usually be present in measures used as both responses and predictors in models such as (1) and more complex models. It is quite rare for these to be taken into account, although they will typically result in biased parameter estimates if ignored. A discussion of the effect of measurement errors is given by Ferrao and Goldstein [21] who show that in one data set the effects of such errors can be large if not properly adjusted for.

Missing data

A standard procedure for handling missing data is via multiple imputation [22] where missingness is assumed to be random, at least conditionally upon other measured variables. In essence this relies upon being able to sample (impute) from a posterior distribution estimated for the value being considered. Missing data values can arise essentially in two ways. One of these is the usual way when values of a measurement, such as a test score, are unobserved. The other way is when an identification, for example of a teacher or school is unknown. The latter case has already been mentioned in the case of a missing teacher identification where it was

suggested that an imputed value for the (unknown) teacher could be obtained according to assumptions regarding the reason for the missing identification. Goldstein [12, Chapter 13] discusses estimation for such models. This is an area that seems to have been little explored and where further empirical data would be useful.

Multivariate models

The extension to multiple outcomes of interest is relatively straightforward, and apart from computational considerations, few new issues occur. The advantage of jointly modelling several outcomes at several levels is that the relationships between different types of institutional effects can be studied and this may be important for certain kinds of judgements.

Computationally, a new issue arises when the outcomes are of different types, such as a mixture of binary, normal and ordered responses. In such cases a 'latent normal' model can be fitted where ordered and binary variables are treated as deriving from underlying normal variables, extending the simple probit analysis model. This can also be extended to unordered categorical variables and count data (Goldstein [12, Chapter 7]). These models therefore allow the joint modelling of data such as exam passes, ordered exam grades and continuously distributed test scores, alongside, for example attitude ratings.

For an example of a simple joint model studying mathematics and English examination results see Goldstein et al. [23].

Modelling where outcomes are measured at higher levels

So far I have described cases where the outcome of interest is measured at the lowest level of the data hierarchy, such as students in schools.

Consider, for example, the case of policing where there is interest in comparing police forces or policing areas in terms of crime rates of different types. While the rate for an area is essentially an aggregation of individually reported incidents, there will often be few if any measurements at the level of the individual incident that are relevant for adjustment purposes. Thus, if we were interested in the efficiency of a police force in tackling burglary, it might be relevant to take account of the vulnerability of the properties where burglaries were reported, since this may differ among areas. Yet often such information may be available only in aggregate form at the area level. We may still use aggregate level variables for adjustment, but in general these will be less efficient, and if there is a small number of areas care will be needed to avoid over-fitting based on a large number of correlated covariates.

In other cases data may only be defined at institutional level. For example rankings of reputation for universities are typically based upon responses from individuals asked to rate institutions

[24]. This case can be treated as one where for each institution we have a number of responses to be aggregated. Unlike the case of school exam results, however, we do not have independent responses across institutions since each rater provides a measure for each university, for example on a simple rating scale. Formally this can be modelled as a cross classification of raters by institutions and a simple model, assuming normality, can be written as

$$y_{\{j_1 j_2\}} = (X\beta)_{\{j_1 j_2\}} + u_{j_1} + u_{j_2} + e_{\{j_1 j_2\}} \qquad (4)$$

$$e_{\{j_1 j_2\}} \sim N(0, \sigma_e^2), \quad u_{j_1} \sim N(0, \sigma_{u1}^2), \quad u_{j_2} \sim N(0, \sigma_{u2}^2)$$

In this model adjustment variables (*X*) may be obtained from the raters or measured at the institutional level. If uncertainty intervals are required for such rankings then they need to take account of the data structure as derived from a model such as (4), and will typically be larger than naïve estimates that treat the responses as independent.

Adjusting for rater characteristics will be especially important since such rankings are often derived using 'convenience' samples such as those derived from databases of journal authors. Thus, for example, larger institutions will tend to produce more authors and therefore have greater representation in the samples used. In this case careful consideration would need to be given to the possibility of weighting respondents to obtain what might be considered a 'representative' sample, although this is clearly a matter for debate. [1]

Examples

We now look at two examples that apply some of the above models and show how the results are relevant within an educational accountability framework.

Our first example is a data set on a cohort of 5748 students in 66 Secondary schools in Inner London. They entered their Secondary (High) schools in year 7 (ages 11 and 12 years) and took school leaving examinations in year 11 (1987). Further details are available in [23]. Cases with any missing data (35%) were excluded from this analysis. A study of these did not suggest any serious biases among those with complete data. Ideally a more efficient analysis could be carried out using multilevel multiple imputation procedures (Goldstein et al., [25]) but for simplicity of illustration we will present only the complete case analysis.

Since Secondary schools differ in the mean academic achievements of the students entering, we employ a 'value added' model where the principal variable used to adjust for intake achievement is a reading test score (London Reading Test, LRT) taken during the year before entry. The basic aim of the analysis is to explore the extent to which schools can be held

[1] Additionally, the samples obtained in such surveys often have response rates as low as about 10% which raises additional concerns about bias. (Phil Baty, personal communication)

accountable for the examination results of their students, after adjusting for selection factors (prior achievement scores) and also looks at differences between curriculum subjects. The response (y) and the LRT score (x_1) are both transformed to have standard normal distributions. Other predictors are gender (x_2 with boys as the reference category), school gender, Girls school (x_3), Boys school (x_4) and mixed gender school (the reference category), and school denomination, Church of England (CE, x_5), Roman catholic (RC, x_6) and State maintained (the reference category). In addition the results of a verbal reasoning test, also taken prior to entry, are used where students are grouped into 3 categories representing approximately 25%, 50% and 25% of the distribution: the three categories are VR1 (x_7), VR2 (x_8) and VR3 as the reference category.

The final fitted model where the normalised examination score is the response is as follows:

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_1^{(2)} x_{1ij}^2 + \beta_2 x_{2ij} + \sum_{k=3}^{4} \beta_k x_{kij} + \sum_{k=5}^{6} \beta_k x_{kij} + \sum_{k=7}^{8} \beta_k x_{kij} + u_{0j} + u_{1j} x_{1ij} + u_{2j} x_{5ij} + e_{ij}$$

$$\begin{pmatrix} u_{0j} \\ u_{1j} \\ u_{2j} \end{pmatrix} \sim N(0, \Omega_u), \qquad e_{ij} \sim N(0, \sigma_0^2 + \sigma_{01} x_{1ij}) \tag{5}$$

Thus, at the school level we have an overall (intercept) effect, a linear component of the relationship with LRT that varies across schools and a difference between the VR1 and combined VR2+VR3 categories, that varies across schools, and it is assumed that these have a 3-variate normal distribution. At the pupil level we assume a normally distributed residual with a variance that is a linear function of the LRT score. Table 1 gives maximum likelihood estimates for this model; see [9, Chapter 2] for details.

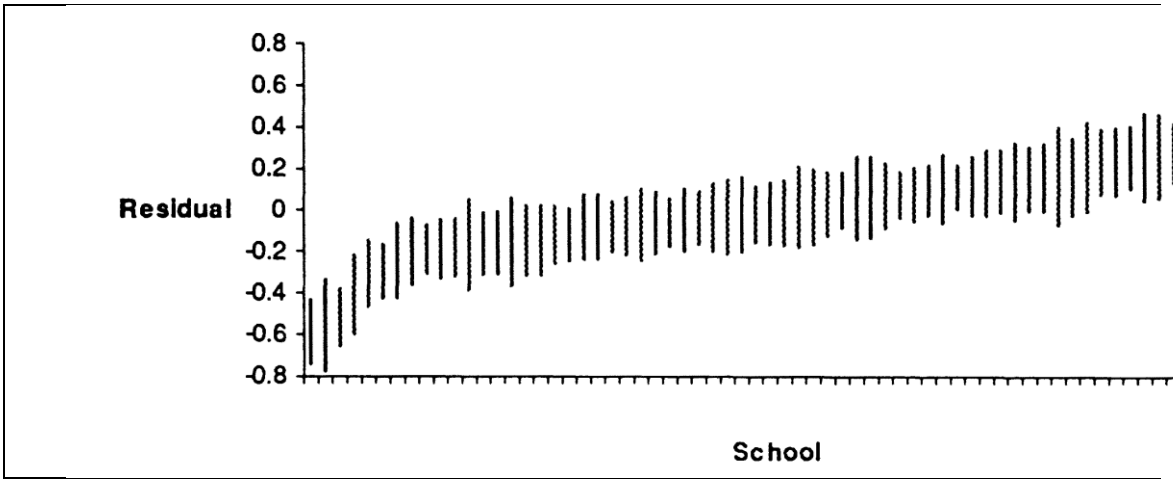
Table 1. Analysis of total examination score.

Fixed coefficients	***Estimate (standard error)***		
β_0	-0.53		
β_1	0.37 (0.02)		
$\beta_1^{(2)}$	0.035 (0.008)		
β_2	0.13 (0.03)		
β_3	0.07 (0.06)		
β_4	0.09 (0.07)		
β_5	-0.04 (0.13)		
β_6	0.20 (0.06)		
β_7	0.70 (0.04)		
β_8	0.31 (0.03)		
Between-school variation $(\boldsymbol{\Omega_u})$. $(\boldsymbol{Correlation})$			
	u_0	u_1	u_2
u_0	0.055		
u_1	0.012 (0.75)	0.0046	
u_2	0.013 (0.40)	0.009 (0.97)	0.019
Between student variation			
σ_0^2	0.55		
σ_{01}	0.046		

For the fixed coefficients we see, as expected, large and statistically significant (at the 5% level) effects for the LRT score with a quadratic relationship indicated, for the VRT category, for gender, with girls on average scoring higher than boys, and for attendance at a Roman Catholic school. At level 2 there is statistically significant variation associated with the LRT score ($\bar{\chi}_3 = 24.7, P < 0.001$) and the VR1 category ($\bar{\chi}_3 = 11.0, P = 0.008$) where in each case the null hypothesis is that the variance term and two associated covariances are zero. The chi-bar test statistic is used [9, Chapter 2] since the variance term is constrained to be non-negative. At level 1 there is a significant positive relationship of the variance with the LRT score ($\chi^2 = 66.0, p < 0.001$).

Using these model estimates we show in figure 2 (corresponding to Figure 1 above) the residual or school effects for the reference categories and the mean value (0) of LRT.

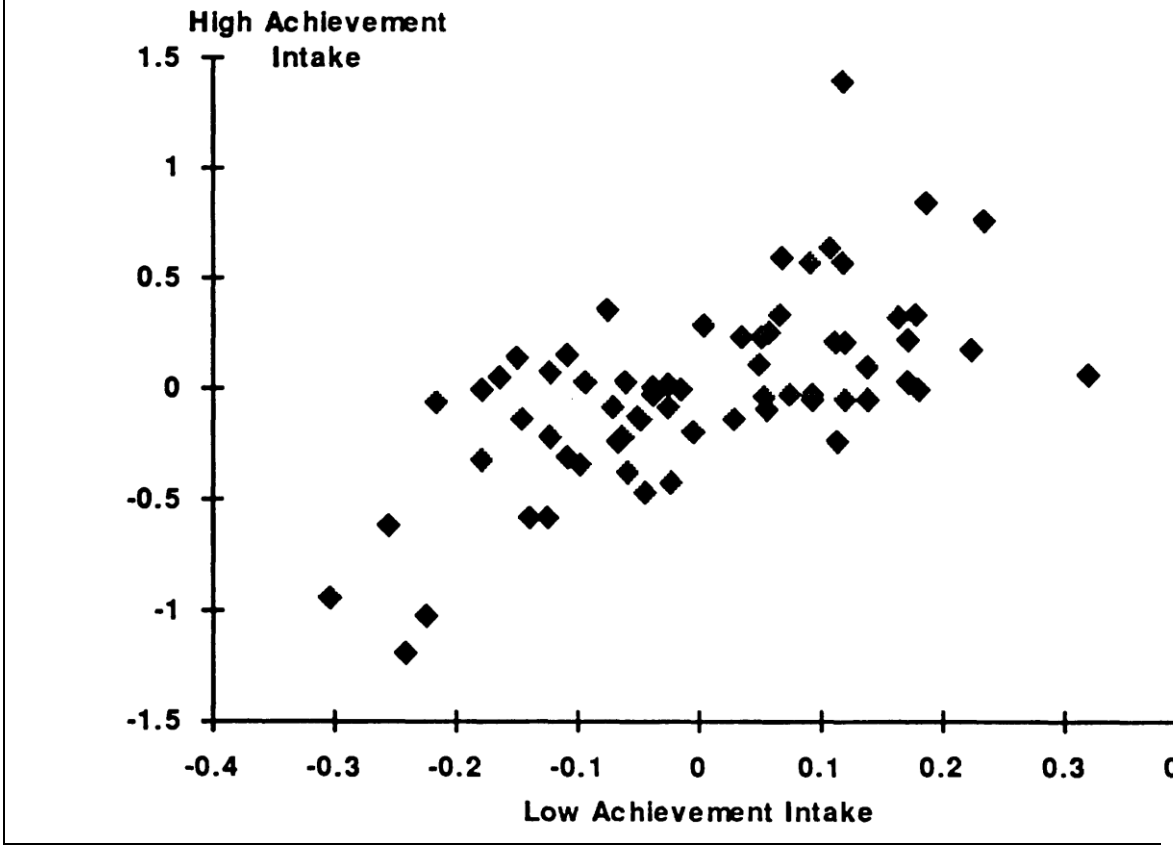
Figure 2. Residual estimates with 95% confidence intervals for examination score data.



We see again the marked uncertainty associated with comparisons among schools.

Since the school effect is also a function of LRT score and VR band we can estimate the effect at different values. Thus, for example we can take two extreme groups, the low achievers at intake with an LRT score of -2 (approximately the lower 2.5 percentile) and in VR bands 2 or 3, compared with the high achievers at intake with an LRT score of 2 (approximately the upper 97.5 percentile) and VR band 1. Figure 3 shows the scatterplot of these estimated from the model as $(u_{0j} - 2u_{1j},\ u_{0j} + 2u_{1j} + u_{2j})$ respectively.

Figure 3. High versus low achievers: school residual estimates

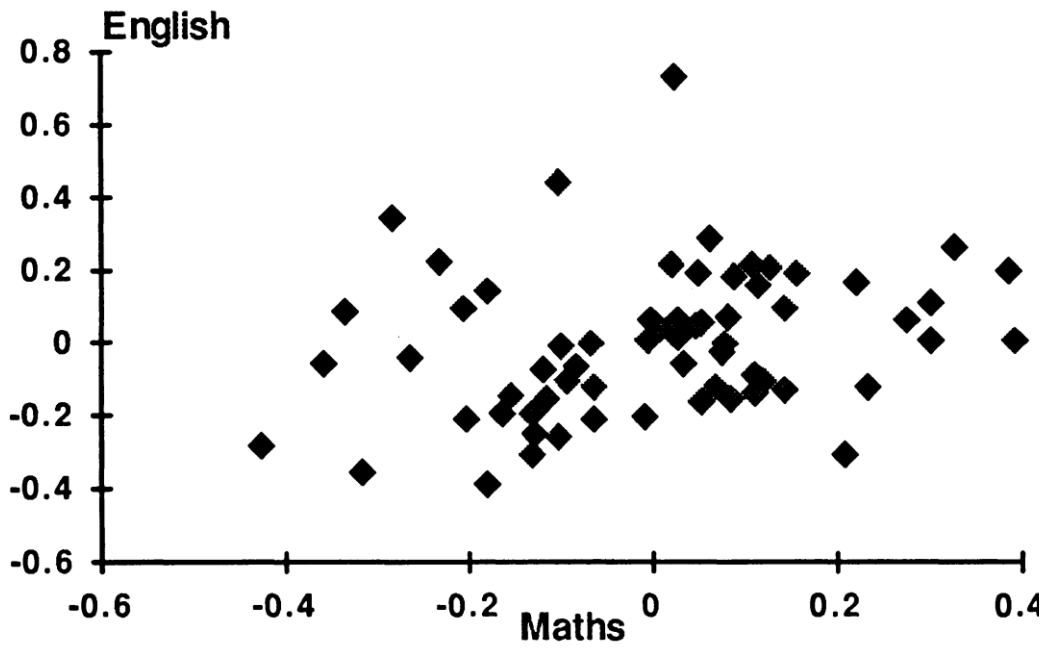


While there is a moderate correlation between these it can be seen that schools do appear to differ in terms of how different types of students perform.

An additional analysis was undertaken [23] in which separate examination scores where English and Mathematics scores were analysed jointly in a bivariate 2-level model. This allows us to estimate a school effect for both mathematics and English and Figure 4 shows the relationship between these estimated residuals.

The estimated correlation between the Mathematics and English school effects is only about 0.1 and this is reflected in Figure 4. Together with the differential effectiveness as a function of intake achievement and the large amounts of uncertainty, any use of such data in the form of overall (unidimensional) league tables will be highly problematic. Nevertheless, for screening purposes, to identify schools that may be performing unexpectedly poorly or well, identifying such schools from such plots may be useful and this is discussed in more detail in [2] and [3]. Our next example looks at school comparisons further in the specific context of school choice.

Figure 4. English and Mathematics examination scores estimated residuals.



This example is a national data set (the National Pupil Database, NPD) which contains longitudinal performance data on all students within maintained (state funded) schools in England. Further details can be found in Leckie and Goldstein, [10,26]. The students are allocated a unique identification when they enter the system and events such as school changes, and test and exam scores are recorded, together with limited demographic information and data on the schools they attend. This database is used both for research purposes and to produce

annual league tables of schools based upon test and examination scores, both unadjusted and adjusted. Within an accountability context one of the uses claimed for these tables is that they assist parents in choice of schools, especially secondary schools and that this process will favour the choice of 'good' schools in terms of how they promote student achievements.

In the context of secondary school choice, a parent who decides to base a choice, at least in part, on such a school ranking will generally have available, at best, results that apply to the previous year's cohort. Their interest, however, is in the future performance of the school in five or six years' time when their child will take the equivalent examination, say in year (grade) eleven. The problem thus becomes a prediction problem from a current set of school effects to a future set of school effects. Leckie and Goldstein [10] utilise GCSE (school leaving) examination data for two cohorts over a six year period (2005 and 2010) where prior achievement data on both cohorts are available.

The relevant model for the two cohorts of students can be written as

$$\begin{aligned} y_{ij}^{(1)} &= \beta_0^{(1)} + \beta_1^{(1)} x_{ij}^{(1)} + u_j^{(1)} + e_{ij}^{(1)} \\ y_{ij}^{(2)} &= \beta_0^{(2)} + \beta_1^{(2)} x_{ij}^{(2)} + u_j^{(2)} + e_{ij}^{(2)} \end{aligned} \qquad (6)$$

$$\begin{bmatrix} u_j^{(1)} \\ u_j^{(2)} \end{bmatrix} \sim N\left(0, \Omega_u\right), \qquad \Omega_u = \begin{bmatrix} \sigma_{u1}^2 & \\ \sigma_{u12} & \sigma_{u2}^2 \end{bmatrix}$$

$$\begin{bmatrix} e_{ij}^{(1)} \\ e_{ij}^{(2)} \end{bmatrix} \sim N\left(0, \Omega_e\right), \qquad \Omega_e = \begin{bmatrix} \sigma_{e1}^2 & \\ 0 & \sigma_{e2}^2 \end{bmatrix}$$

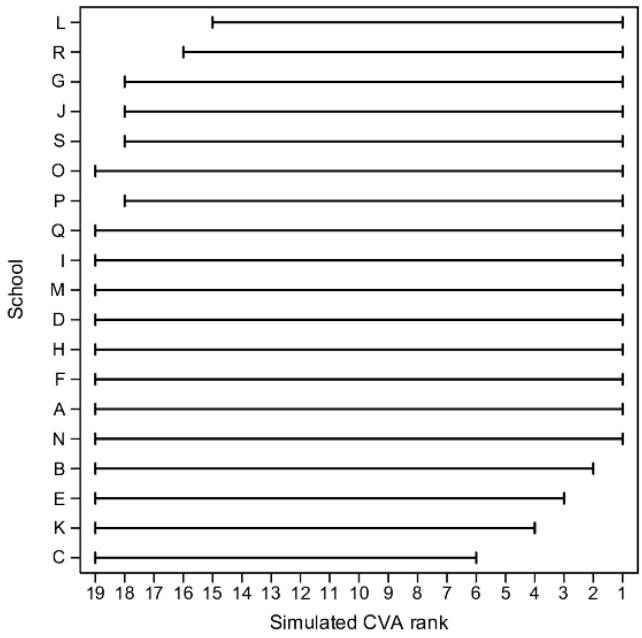
where superscripts '(1)' and '(2)' denote cohort 1 and cohort 2. Hence $y_{ij}^{(1)}$ is the GCSE score for the *i*th pupil in the *j*th school in cohort 1 (2005) whilst $y_{ij}^{(2)}$ is the GCSE score for the *i*th pupil in the *j*th school in cohort 2 (2010). The level 2 school residuals in general will be correlated. The level 1 residuals for the two responses are modelled as independent as a pupil can only belong to one cohort. Hence, this is a bivariate model where the bivariate structure is at level 2 rather than in the traditional multivariate multilevel model where it is at both levels.

From this model, having obtained the parameter estimates we can obtain estimates of the cohort 2 predicted school effects as functions of the terms in Ω_u, Ω_e and the $y_{ij}^{(1)}$. For school *j* this is given by

$$\frac{\rho_{u12} n_j^{(1)} \sigma_u^2}{(n_j^{(1)} \sigma_u^2 + \sigma_{e1}^2)} \tilde{y}_j^{(1)}$$

where $\tilde{y}_j^{(1)}$ is the mean of the raw residuals for the *j*th school in cohort 1, with a corresponding term for the estimated standard error. From this we can construct a caterpillar plot that ranks the predicted school effects together with interval estimates. This is illustrated in Figure 5 that uses the results for 19 Secondary schools in Bristol with 96% confidence intervals. The results are striking with all intervals overlapping, thus providing no reliable separation.

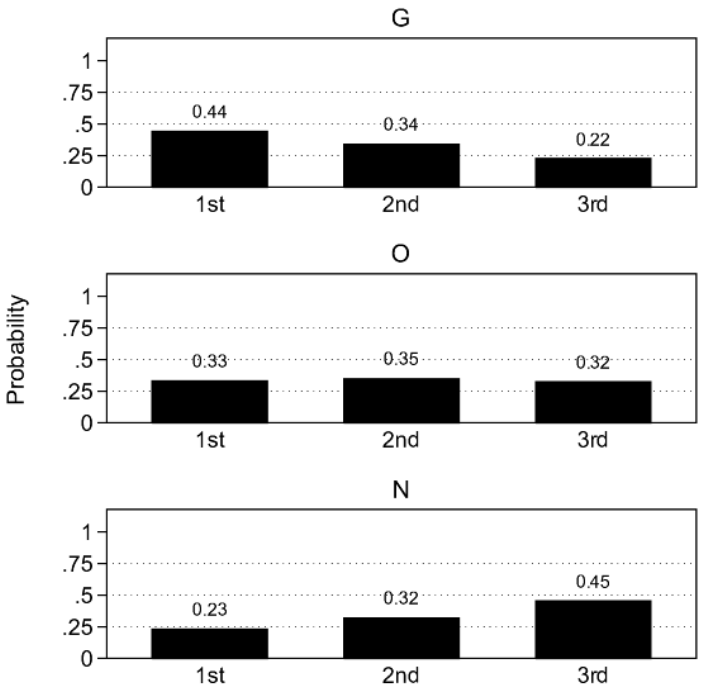
Figure 5. Predicted school effects for 19 Bristol schools with 95% confidence intervals.



Leckie and Goldstein[26] go on to present school comparisons in terms of probabilities, that for any given set of schools forming a choice set, any particular one will have the largest or smallest predicted effect. Figure 6 illustrates this for three chosen schools.

For none of the schools is there a better than even chance of being ranked first. Displays such as this can convey the extensive uncertainty in ways that are clearly intelligible to non-technical audiences.

Figure 6 Probability that school G, N and O are predicted to be ranked 1st, 2nd or 3rd



Conclusions

As pointed out in the introduction, I have concentrated on a consideration of the statistical issues, with illustrative examples, and I have not spent time discussing side effects including 'perverse incentives' for institutions to behave in ways that may not serve the best interests of those whom they are meant to serve, be they students, patients or the general public. All of these issues are, however, both important and researchable and the producers of league tables need to do more to encourage such research. A fuller discussion is given in [3]. The evidence, where suitable data are available, is that rankings of institutions have large measures of uncertainty attached to them, even when appropriate adjustments for selection effects have been made. Perhaps the most effective uses of institutional rankings are as screening instruments that can suggest where problems may be occurring, rather than diagnoses of what the problems are.

I am not suggesting that league tables should never be published. There is clearly a need for accountability from public (and other) institutions and quantitative data that bear on performance are a useful tool for this. When such data are reported publicly, however, their quality and reliability need to be displayed also so that users of the data are not misled about what can be inferred. To withhold information about the uncertainty of rankings is to deprive users of information they are entitled to.

While there may be useful work to be done in the further development the models described here, a more pressing need is to find ways of enhancing data quality and especially ways of preventing unrealistic inferences being drawn from over-simple presentations of results. For example, it is perfectly possible to develop software for sites that host institutional databases, in order to provide information similar to that in Figure 6. This could be done in real time and display bespoke comparisons among institutions. By displaying the real uncertainty surrounding institutional comparisons it would help users to make properly informed judgements.

References

1. Hood, C. (2007), "Public Service Management by Numbers: Why does it vary? Where has it come from? What are the gaps and puzzles?" in Public Money and Management, Vol. 27, No.2, 95-102
2. Yang M, Goldstein H, Rath T, Hill N. (1999). The Use of Assessment Data for School Improvement Purposes, . Oxford Review of Education, Vol. 25, No. 4 pp.469-483
3. Foley, B., and Goldstein, H. (2012). *Measuring Success: league tables in the public sector*. London, British Academy.
4. Normand, S.T. (2014). League tables in hospital comparisons. Annual review of statistics and its applications, in……
5. Dill, D., and Soo, M. (2005) "Academic Quality, League Tables and Public Policy: A cross-national analysis of university ranking systems" in Higher Education, 49(4), 495-537
6. Lockwood, J. R., McCaffrey, D. F., Mariano, L. T., & Setodji, C. (2007) Bayesian methods for scalable value-added assessment. Journal of Educational and Behavioral Statistics. Vol 32(2), 125-150.
7. Bird, S., et al. (2005). Performance indicators: good bad, and ugly. Journal of the Royal Statistical Society, A. **168**: 1-27.
8. Goldstein, H. and Spiegelhalter, D. J. (1996). League tables and their limitations: statistical issues in comparisons of institutional performance. Journal of the Royal Statistical Society, A. **159**: 385-443
9. Goldstein, H. (2011). Multilevel Statistical Models. Fourth edition. Chichester, Wiley.
10. Leckie, G. and H. Goldstein (2009). "The limitations of using school league tables to inform school choice." Journal of the Royal Statistical Society, A **172**, 835-851

11. Goldstein H & Healy M J R. (1995). The Graphical Presentation of a Collection of Means.. Journal of the Royal Statistical Society, 581, 1, pp 175-177.
12. Hazelkorn, E. (2011). *Rankings and the reshaping of higher education: the battle for world-class excellence.* Palgrave MacMillan. ISBN 978-0-230-24324-8
13. Spiegelhalter, D. J. and Goldstein, H. (2009). "Comment: Citation Statistics." Statistical Science , **24**, 21-24
14. Goldstein H & Sammons P.. (1997). The Influence of Secondary and Junior Schools on Sixteen Year Examination Performance: A Cross-classified Multilevel Analysis. School Effectiveness and School Improvement, Vol. 8. No. 2, pp. 219-230.
15. Goldstein, H., Burgess, S., and McConell, B. 2007. Modelling the effect of pupil mobility on school differences in educational achievement. Journal of the Royal Statistical Society, A, 170, 4, 941-954
16. Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and Van der Linde, A. (2002). Bayesian measures of model complexity and fit. Journal of the Royal statistical Society, B, 64, 583-640.
17. Nuttall, D.L., Goldstein,H., Prosser,R. and Rasbash,J. (1989). Differential school effectiveness. International Journal of Educational Research, 13, 769-76.
18. Yang M, Goldstein H, Rath T, Hill N. (1999). The Use of Assessment Data for School Improvement Purposes, . Oxford Review of Education, Vol. 25, 4, pp.469-483.
19. Hubbard, A.E., Ahern, J., Fleischer, N.L., Van der Laan, M., Lippman, S.A., Jewell, N., Bruckner, T. and Satariana, W.A. (2010). To GEE or Not to GEE. *Epidemiology,* 21, 4, 467-474.
20. Amrein-Beardsley, A., Collins, C., Polasky, SA., and Sloat, EF. (2013). Value-Added Model (VAM) research for Educational Policy: Framing the Issue. Education policy analysis archive; **21,4.** January 28th, 2013.
21. Ferrao, M. and Goldstein, H. (2012). Adjusting for differential misclassification in multilevel models: the relationship between child exposure to smoke and cognitive development. Quality and quantity. Published online; DOI 10.1007/s11135-012-9765-5.
22. Rubin, D. (1987). Multiple imputation for non-response in surveys. Chichester: Wiley.
23. Goldstein H, Rasbash J, Yang M, Woodhouse G, Pan H, Nuttall, D. & Thomas S. A. (1993). Multilevel analysis of school examination results. Oxford Review of Education **19** (4) 425-433.
24. Baty, P. (2010) "Measured, and found wanting more" in Times Higher Education, 8th July 2010.
25. Goldstein, H., Carpenter, J., Kenward, M. and Levin, K. (2009). Multilevel models with multivariate mixed response types. Statistical Modelling, 9,3, 173-197.

26. Leckie, G. and Goldstein, H. (2011). Understanding uncertainty in school league tables. Fiscal Stdies, 32, 207-224.